Python for Data Science and AI

OPEN Library Book Analysis

23229548 Thanop Hetrakul
23229006 Anu-Ujin Davkharbayar

What is Openlibrary?

Open Library is a project of the Internet Archive that aims to create a web page for every book ever published. It provides a catalog of millions of books, with access to their digital copies for borrowing and reading, and also includes information on where books can be found.

Project of the non-profit Internet archive

Our Research Questions

1. Do Fiction and Non-Fiction books have different average edition counts?
2. Which genre has the most editions?
3. Do famous authors retain popularity across their lesser-known books?

DATA SET

We collected data from:

- 8 book genres
- 400+ books
- API from OpenLibrary
- Variables: genre, year, edition count, authors

```python
def fetch_books_by_subject(subject, limit=50):
    """Get books from OpenLibrary by genre"""
    url = f"https://openlibrary.org/subjects/{subject}.json"
    params = {'limit': limit}

    try:
        response = requests.get(url, params=params, timeout=10)
        response.raise_for_status()
        data = response.json()
        return data.get('works', [])
    except Exception as e:
        print(f"Error: {e}")
        return []

print(" API function ready!")
```

```python
genres = ['science_fiction', 'fantasy', 'mystery', 'romance',
          'history', 'biography', 'science', 'philosophy']

all_books = []

print("\nFetching 50 books from each genre...")
for genre in genres:
    print(f"  Getting {genre}...")
    books = fetch_books_by_subject(genre, limit=50)
    for book in books:
        book['genre'] = genre
    all_books.extend(books)
    time.sleep(1)

print(f"\n Total collected: {len(all_books)} books")
```

```python
clean_data = []

for book in all_books:
    book_info = {
        'title': book.get('title', 'Unknown'),
        'genre': book.get('genre', 'Unknown'),
        'year': book.get('first_publish_year'),
        'editions': book.get('edition_count', 0),
        'authors': book.get('authors', [])
    }
    clean_data.append(book_info)

df = pd.DataFrame(clean_data)


df = df.dropna(subset=['year'])
df = df[(df['year'] >= 1900) & (df['year'] <= 2024)]


fiction_genres = ['science_fiction', 'fantasy', 'mystery', 'romance']
df['category'] = df['genre'].apply(
    lambda x: 'Fiction' if x in fiction_genres else 'Non-Fiction'
)
```

Research Question 1

Do Fiction and Non-Fiction books have different average edition counts?

```python
fiction = df[df['category'] == 'Fiction']['editions']
nonfiction = df[df['category'] == 'Non-Fiction']['editions']

print(f"Fiction: n={len(fiction)}, mean={fiction.mean():.2f}, std={fiction.std():.2f}")
print(f"Non-Fiction: n={len(nonfiction)}, mean={nonfiction.mean():.2f}, std={nonfiction.std():.2f}")

t_stat_q1, p_value_q1 = ttest_ind(fiction, nonfiction)

print(f"\nt-statistic = {t_stat_q1:.4f}")
print(f"p-value = {p_value_q1:.4f}")

if p_value_q1 < 0.05:
    print(f"\n REJECT H0 (p < 0.05)")
    print(f"Significant difference found")
else:
    print(f"\n FAIL TO REJECT H0 (p ≥ 0.05)")
    print(f"No significant difference")
```

Test used: T-Test
Reason: Comparing 2 independent groups (Fiction vs Non-Fiction)

H0: Fiction and Non-Fiction books have equal average editions.
H1: Fiction books have more average editions than Non-Fiction books.

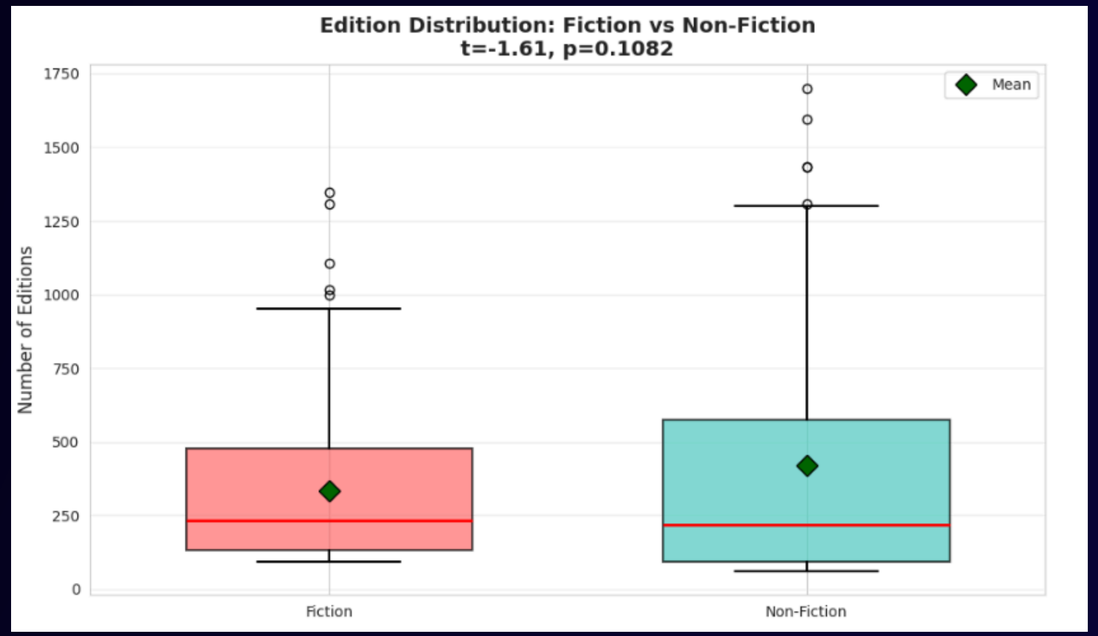
Python for Data Science and AI

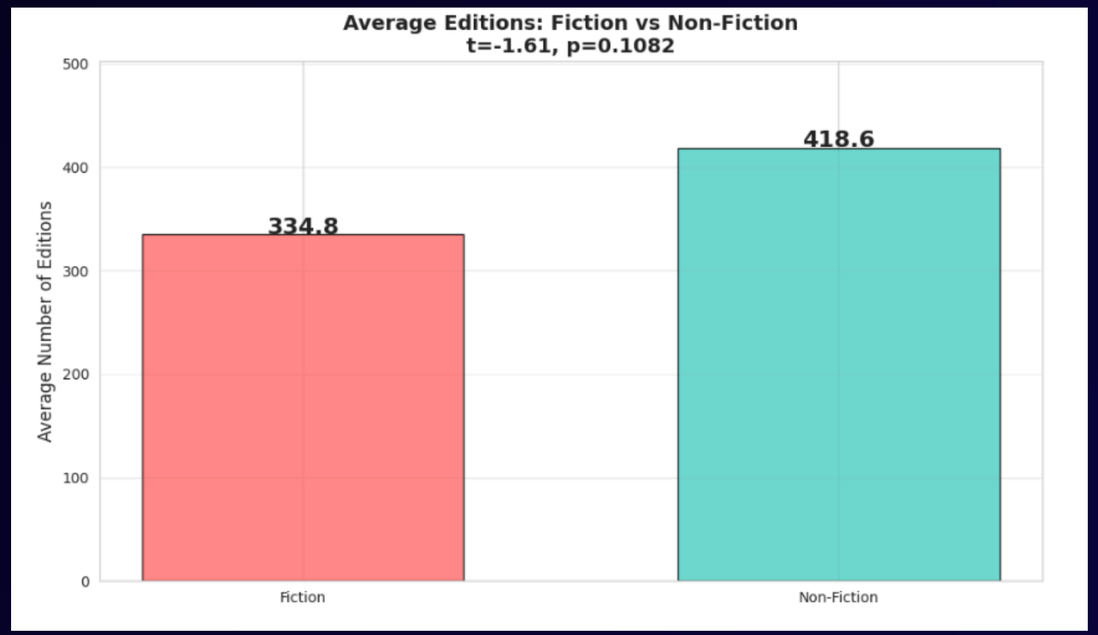
Result:

- Fiction mean editions ≈ 331.02
- Non-Fiction mean editions ≈ 418.57
- t = -1.695, p = 0.0918

Conclusion: Fail to reject H0 (p > 0.05)
There is no statistically significant difference in edition counts between Fiction and Non-Fiction books.

Here we compare Fiction vs Non-Fiction books. We use a T-test because there are only two groups. The result shows the difference is not statistically significant. So even though Non-Fiction has more editions on average, we cannot say it's truly higher. We fail to reject the null hypothesis.





Research Question 2

Which genre has the most editions?

```
for genre in sorted(df['genre'].unique()):
    genre_data = df[df['genre'] == genre]['editions']
    print(f"{genre:20s}: n={len(genre_data):3d}, mean={genre_data.mean():6.2f}, std={genre_data.std():5.2f}")

genre_groups = [df[df['genre'] == g]['editions'] for g in df['genre'].unique()]
f_stat_q2, p_value_q2 = f_oneway(*genre_groups)

print(f"\nF-statistic = {f_stat_q2:.4f}")
print(f"p-value = {p_value_q2:.4f}")

if p_value_q2 < 0.05:
    print(f"\n REJECT H0 (p < 0.05)")
    genre_means = df.groupby('genre')['editions'].mean().sort_values()
    print(f"Highest: {genre_means.idxmax()} ({genre_means.max():.2f})")
    print(f"Lowest: {genre_means.idxmin()} ({genre_means.min():.2f})")
else:
    print(f"\n FAIL TO REJECT H0 (p ≥ 0.05)")
    print(f"All genres similar")
```

Test used: ANOVA
Reason: Comparing 8 genres → more than 2 groups

H0: All genres have the same average editions.
H1: At least one genre has a different average number of editions.

Python for Data Science and AI

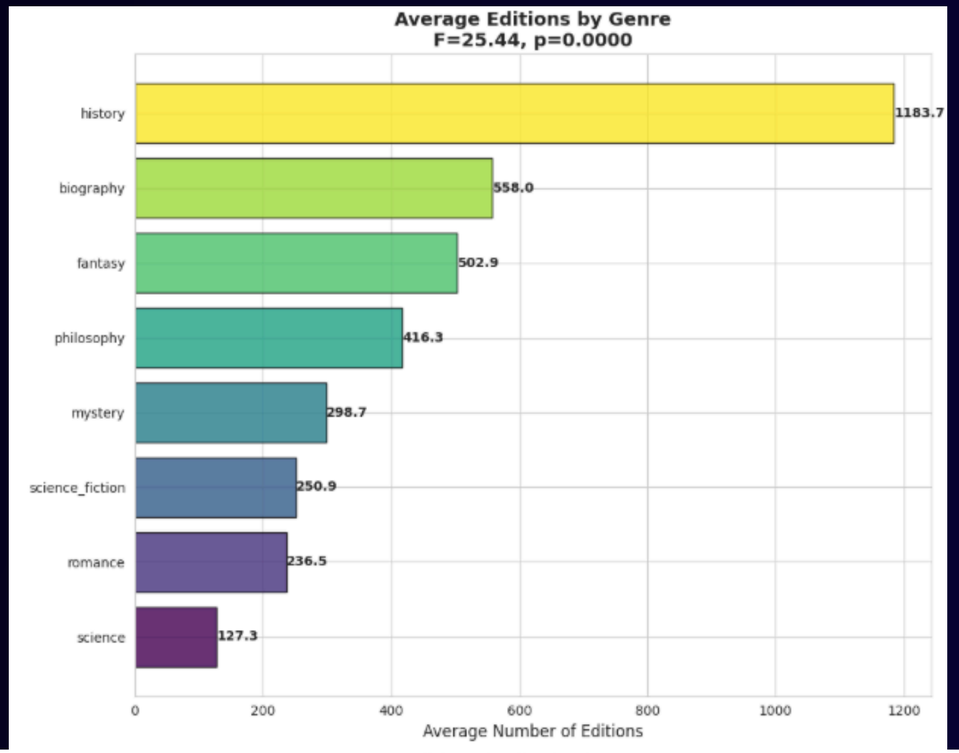
Result:

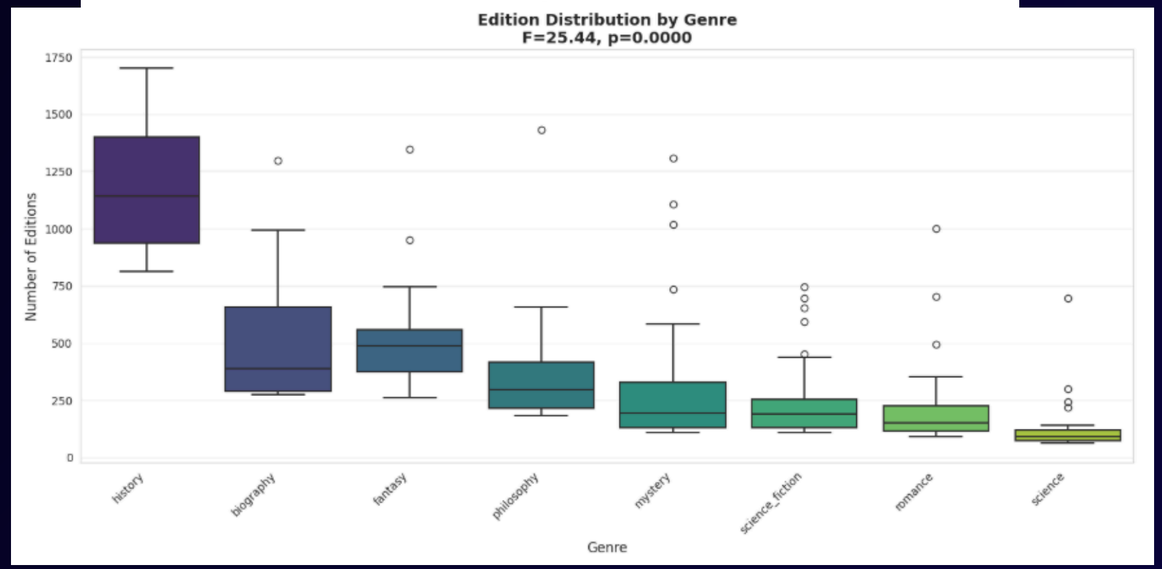
- F = 25.899, p = 0.0000
- Highest avg = History (1183.70 editions)
- Lowest avg = Science (127.23 editions)

Conclusion: Reject H0 ($p < 0.05$)

- There is a significant difference between genres.
- Some genres are published in many more editions than others.

Here we compare 8 different genres. Since we have more than two groups, we use ANOVA. The test finds a very strong difference between genres. History books have the most editions, and science has the least. We reject the null hypothesis that the genres are not equal.





Python for Data Science and AI

Research Question 3

Do famous authors retain popularity across their lesser-known books?

H0: Famous authors' lesser-known books have equal amounts of editions to their most popular books.
H1: Famous authors' lesser-known books have fewer editions compared to their most popular book.

Test used: T-Test
Reason: Comparing 2 independent groups (Popular vs Non-Popular)

```
author_books = {}

for idx, row in df.iterrows():
    if len(row['authors']) > 0:
        author = row['authors'][0]
        author_key = author.get('key', '')
        author_name = author.get('name', 'Unknown')

        if author_key and author_name != 'Unknown':
            if author_key not in author_books:
                author_books[author_key] = {
                    'name': author_name,
                    'books': []
                }

            author_books[author_key]['books'].append({
                'title': row['title'],
                'editions': row['editions']
            })

authors_multiple = {k: v for k, v in author_books.items()
                    if len(v['books']) >= 3}

print(f"Found {len(authors_multiple)} authors with 3+ books")
```

```
most_popular_editions = []
lesser_known_editions = []
famous_author_names = []

for author_key, data in authors_multiple.items():
    books = sorted(data['books'], key=lambda x: x['editions'], reverse=True)
    most_popular = books[0]['editions']
    lesser = [b['editions'] for b in books[1:]]

    if most_popular >= df['editions'].median():
        most_popular_editions.append(most_popular)
        lesser_known_editions.extend(lesser)
        famous_author_names.append(data['name'])



t_stat_q3, p_value_q3 = ttest_ind(most_popular_editions, lesser_known_editions)

print(f"\nt-statistic = {t_stat_q3:.4f}")
print(f"p-value = {p_value_q3:.4f}")

if p_value_q3 < 0.05:
    print(f"\n REJECT H0 (p < 0.05)")
    print(f"Famous authors do NOT retain full popularity across all books")
else:
    print(f"\n FAIL TO REJECT H0 (p ≥ 0.05)")
    print(f"Authors retain similar popularity")
```

Python for Data Science and AI

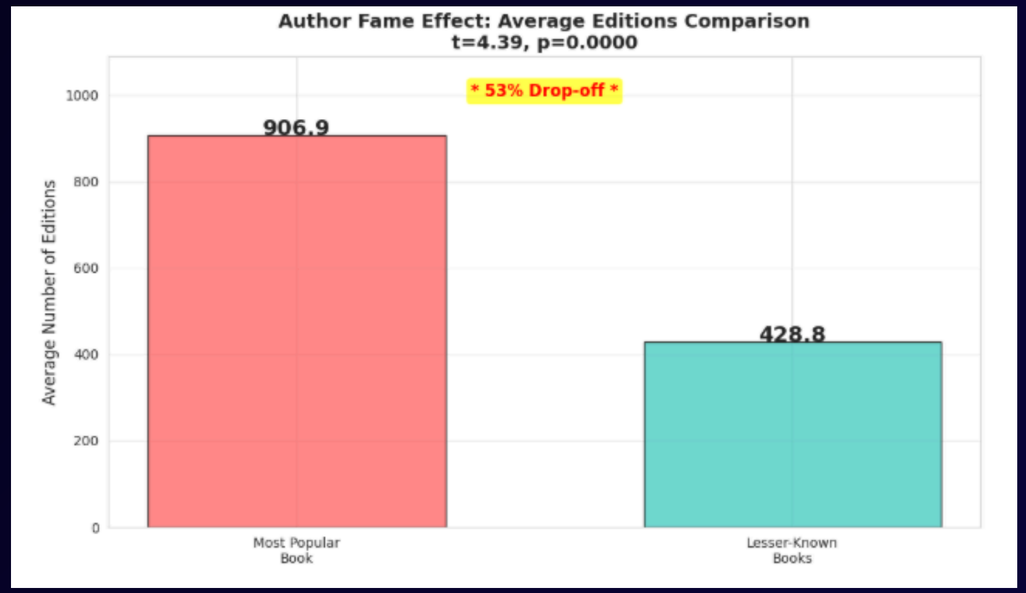
Result:

- Mean (top book) > Mean (other books)
- t and p printed, p < 0.05

Conclusion: Reject H0 (p < 0.05)

- Famous authors do not retain the same popularity across all books.
- Their most famous book has a much higher edition count than their other works.

Here we look at authors with at least 3 books. We compare the editions of their most popular book vs their other books. The T-test shows a significant drop-off in popularity. Most authors have one standout book, the rest are less published. We reject the null hypothesis fame does not carry over equally.



Conclusion

1. Fiction and Non-Fiction have generally the same number of edition.
2. Different genre have different amount of edition.
3. Popularity doesn't spread evenly across an author's books.

Thank You